IE7374 FINAL PROJECT REPORT
Air Quality Index Prediction

Group 9

Ayush Soni
Divyangana Pandey
Radhika Jayachandran
Snehal Rajwar

soni.ay@northeastern.edu
pandey.di@northeastern.edu
jayachandran.r@northeastern.edu
rajwar.s@northeastern.edu

Percentage of Effort Contributed by Student 1: 25%
Percentage of Effort Contributed by Student 2: 25%
Percentage of Effort Contributed by Student 3: 25%
Percentage of Effort Contributed by Student 4: 25%

Signature of Student: *Ayush Soni*
Signature of Student: *Divyangana Pandey*
Signature of Student: *Radhika Jayachandran*
Signature of Student: *Snehal Rajwar*

Date:08/08/2022

I. Introduction:

Development in industrialization and advancement of technology has led to an increase in pollutant levels in the air. In China, this has led to an increase in extreme weather conditions such as a higher recorded increase in temperature, haze, smog and adverse effects on health conditions to the residents of these areas.

The research on the concentration of these pollutants will provide the information required for the improvement of significant health, environmental conditions, and benefits to the economy while reducing costs on solutions to reduce air pollution.

Fig 1.1: Image recorded of the same location in Beijing on two separate dates showing differences in air quality

The Air Quality Index is a metric that quantitatively describes the number of pollutants suspended in the air to briefly describe the air quality in an area. The air's concentration of pollutants such as PM2.5, PM10, CO, O3, SO2 and NO2 indicates the overall air quality. However, other meteorological conditions also play a pivotal role in the quality of air.

I.I. Problem Definition:

Through this project, we aim to research and incorporate complex nonlinear relationships between the concentration of air pollutants and meteorological variables across a given period. Implement and build a prediction system based on individual pollutant concentration levels that can predict air quality. Doing so will make the information on the air quality index more

flexible and useful. Systems that can generate warnings based on air quality are required and important for a change as they may play an important role in health alerts when air pollution levels may exceed specified levels. By the end of this classroom project, we aim to establish a relationship between the concentration of pollutants and its correlation with AQI (Air Quality Index).

AQI	Air Pollution Level	Health Implications	Cautionary Statement (for PM2.5)
0 - 50	Good	Air quality is considered satisfactory, and air pollution poses little or no risk	None
51 -100	Moderate	Air quality is acceptable; however, for some pollutants there may be a moderate health concern for a very small number of people who are unusually sensitive to air pollution.	Active children and adults, and people with respiratory disease, such as asthma, should limit prolonged outdoor exertion.
101-150	Unhealthy for Sensitive Groups	Members of sensitive groups may experience health effects. The general public is not likely to be affected.	Active children and adults, and people with respiratory disease, such as asthma, should limit prolonged outdoor exertion.
151-200	Unhealthy	Everyone may begin to experience health effects; members of sensitive groups may experience more serious health effects	Active children and adults, and people with respiratory disease, such as asthma, should avoid prolonged outdoor exertion; everyone else, especially children, should limit prolonged outdoor exertion
201-300	Very Unhealthy	Health warnings of emergency conditions. The entire population is more likely to be affected.	Active children and adults, and people with respiratory disease, such as asthma, should avoid all outdoor exertion; everyone else, especially children, should limit outdoor exertion.
300+	Hazardous	Health alert: everyone may experience more serious health effects	Everyone should avoid all outdoor exertion

Fig 1.2: Matrix showing the AQI levels and their implications on health

II. Data Description:

The dataset was procured from the very versatile website for the Data Science dataset, UCI ML. The dataset records hourly pollutant concentrations from twelve different controlled air quality monitoring centres across China.

II.I. Data Features:

The data set selected by us comprises of eighteen dimensions and four lakh instances across the timeline of March 1st, 2013, to February 28th, 2017. Below we listed the types of features to gauge a better understanding of our data:

Feature	Description
Serial No	Number of Records
Year	Year of data recorded
Month	Month of data recorded
Day	Day of data recorded
Hour	Hour of data recorded
PM 2.5	Particulate Matter 2.5 concentration (ug/m^3) - Tiny particles or droplets in the air that are two and one-half microns or less in width
PM 10	PM10 concentration (ug/m^3) - any particulate matter in the air with a diameter of 10 micrometers or less, including smoke, dust, soot, salts, acids, and metals.
SO2	Sulfur Dioxide concentration (ug/m^3) - colorless gas or liquid with a strong, choking odour and produced from the burning of fossil fuels (coal and oil) and the smelting of mineral ores (aluminum, copper, zinc, lead, and iron) that contain sulfur.
NO2	Nitrogen Dioxide concentration (ug/m^3) - group of highly reactive gases.
CO	Carbon Monoxide concentration (ug/m^3) - odorless, colorless gas formed by the incomplete combustion of fuels.
O3	Ozone concentration (ug/m^3) - highly reactive gas composed of three oxygen atoms.
Temperature	Indicates temperature in °C
Pressure	Indicates atmospheric pressure in Pascals (hPa)
Dew Point	Indicates dew point temperature in °C
Rain	Precipitation (mm)
Wind Direction	Direction of the wind
Wind Speed	Wind speed (m/s)
Station	Name of the air-quality monitoring site

II.II. Data Cleaning and Preprocessing:

To prepare the data for exploration and model fitting, we first combined the data files that were separated by record for each unique station. After combining these files we performed various exploratory data analyses to study and understand the recorded data and the following were the changes made to attain our target:

- Columns ‘year’, ‘month’, ‘day’ and ‘hour’ were combined to create a Date – Time column such that we can distinguish the records per each day per time.
- There were no duplicate records of Date – Time in the dataset, hence no rows were dropped.

II.III. Feature Engineering

The recorded data for these pollutants were converted to their standard unit from the recorded unit. The above calculations were achieved by the formulae:

Pollutant	Recorded Unit	Standard Unit
Temperature	Degree Celsius	Kelvin
PM 2.5	µg/m^3	µg/m^3
PM 10	µg/m^3	µg/m^3
SO2	µg/m^3	Parts per billion
NO2	µg/m^3	Parts per billion
O3	µg/m^3	Parts per million
CO	µg/m^3	Parts per million

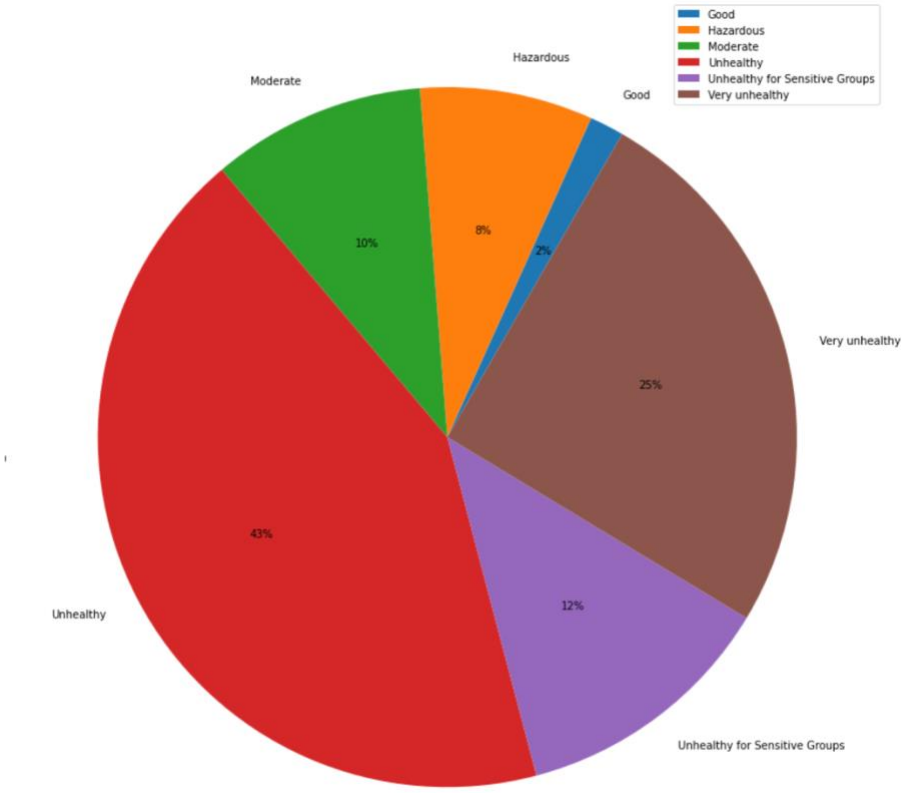
- All Null values within the data were replaced by their median for each day recorded by each station. While checking the data we noted that there exist days with a whole 24-hour period where no data was recorded. We replaced these null values with 0.
- Further to obtain the AQI Index of individual pollutants, a rolling window average for each pollutant must be obtained. This data is then segregated into various classifications of AQI and the max value is recorded as the quality index for the given day. The category of AQI is defined as per the standardized government information.
- AQI is calculated using the formula given below:

$$IAQI_m = \frac{IAQI_{Hi} - IAQI_{Lo}}{BP_{Hi} - BP_{Lo}} (C_m - BP_{Lo}) + IAQI_{Lo}$$

O_3 (ppm) 8-hour	O_3 (ppm) 1-hour[1]	$PM_{2.5}$ (µg/m³) 24-hour	PM_{10} (µg/m³) 24-hour	CO (ppm) 8-hour	SO_2 (ppb) 1-hour	NO_2 (ppb) 1-hour	AQI	
0.000 - 0.054	-	0.0 – 12.0	0 - 54	0.0 - 4.4	0 - 35	0 - 53	0 - 50	Good
0.055 - 0.070	-	12.1 – 35.4	55 - 154	4.5 - 9.4	36 - 75	54 - 100	51 - 100	Moderate
0.071 - 0.085	0.125 - 0.164	35.5 – 55.4	155 - 254	9.5 - 12.4	76 - 185	101 - 360	101 - 150	Unhealthy for Sensitive Groups
0.086 - 0.105	0.165 - 0.204	(55.5 - 150.4)[3]	255 - 354	12.5 - 15.4	(186 - 304)[4]	361 - 649	151 - 200	Unhealthy
0.106 - 0.200	0.205 - 0.404	(150.5 - (250.4)[3]	355 - 424	15.5 - 30.4	(305 - 604)[4]	650 - 1249	201 - 300	Very unhealthy
([2])	0.405 - 0.504	(250.5 - (350.4)[3]	425 - 504	30.5 - 40.4	(605 - 804)[4]	1250 - 1649	301 - 400	Hazardous
([2])	0.505 - 0.604	(350.5 - 500.4)[3]	505 - 604	40.5 - 50.4	(805 - 1004)[4]	1650 - 2049	401 - 500	Hazardous

II.IV. Data Exploration:

Pie Chart - Percentage of Days Within Each Air Quality Classification:



Python has functional visualization libraries that can explore and project your data in creative and pictorial ways. Here we plotted a pie chart of the count of different AQI indexes to show where the air quality stands within the region.

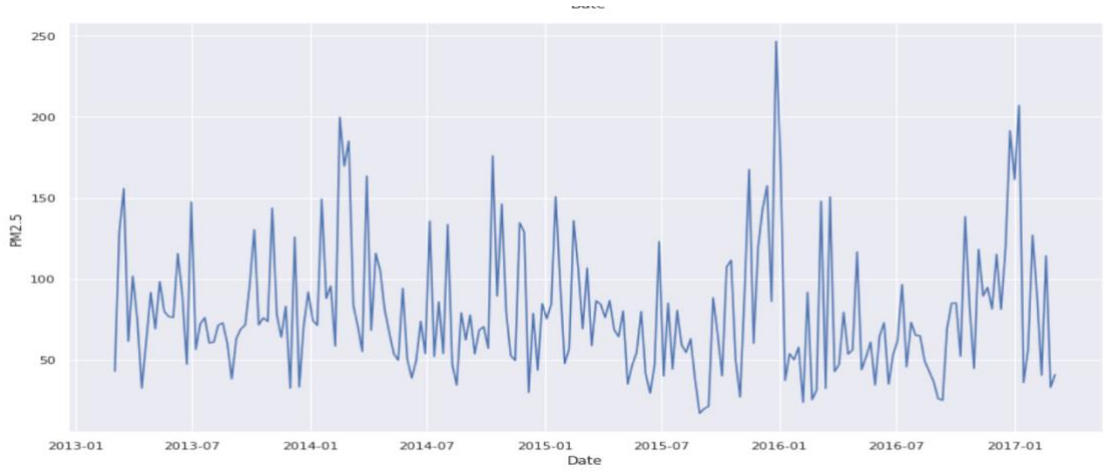
Box Plot – Air Quality Index Across Months:

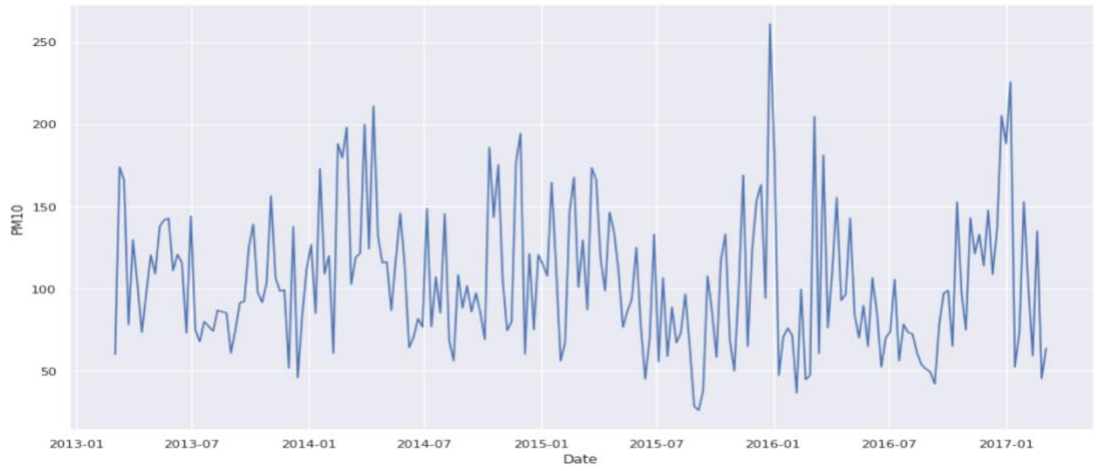
The boxplot gives us a picture of the ranges of AQI for each month to get an idea of which months have the best air quality and the average air quality across each month.

Heat Map – Correlation Matrix Between Input Features:

As we can see above the heat map gives us the level of correlation between our input features from the dataset. This can help remove features that are highly correlated.

Time series – Pollutant Concentration Across Time:

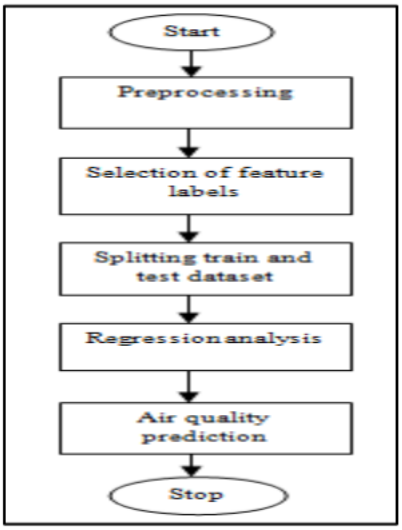




The time series gives us the effect of each pollutant and its curve on air quality over the years.

III. Model Implementation:

We processed and analyzed our data through regression models to obtain accurate predictions. Machine learning models use techniques that find the relationship between dependent and independent variables. These techniques can be used in forecasting, prediction, time series modelling etc. Some of the modelling techniques we learned in the classroom setting that can be applied accurately in our AQI prediction project are Linear regression, Logistic regression and Naïve Bayes.



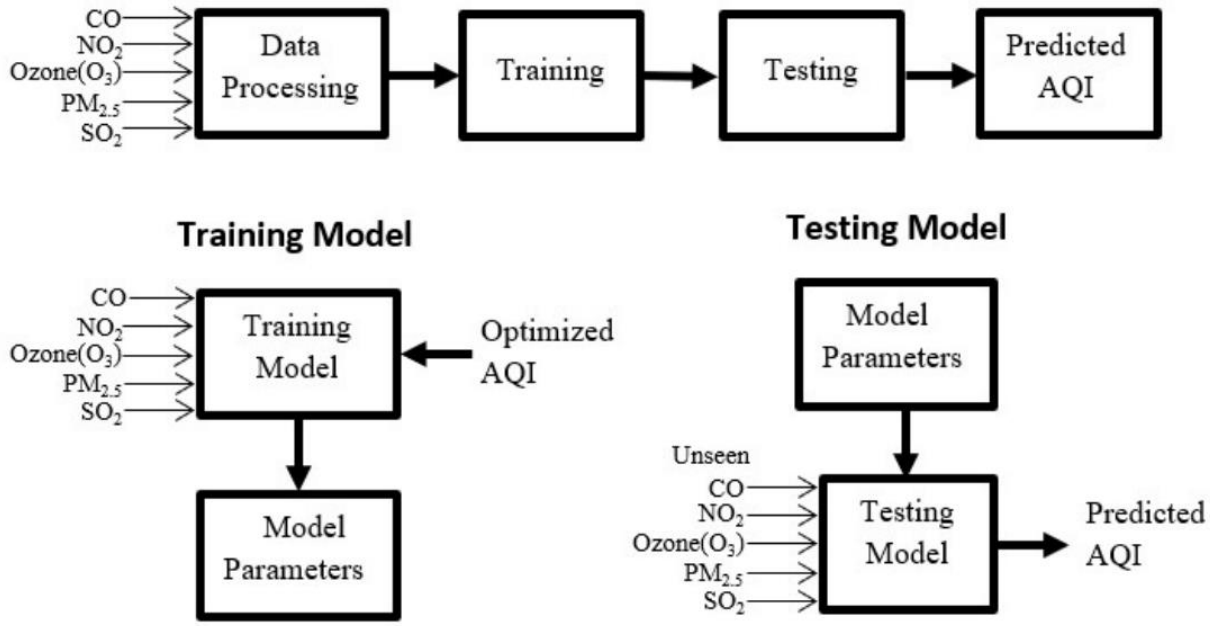
III.I. Linear Regression:

Linear regression is a method of fitting independent variables with dependent variables into a line in n dimensions where n is the total number of features or variables within our data. The model also minimizes the total error while trying to fit all the data points in one line and learning continuously by optimizing parameters. Optimization is carried out by using gradient descent. Gradient descent partially derives the loss function and all parameters are updated by subtracting the previous value with the derivative time the learning rate. The learning rate is tuned by the trial and error method. The equation of multivariate linear regression is given by:

$$Y_i = \alpha + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \ldots + \beta_n x_i^{(n)}$$

Where,

Yi = estimate of ith component of dependent variable y
n = independent variables



Results:

The linear regression model was run for various parameters in order to obtain the most accurate results.

RMSE	Learning Rate	Tolerance	Maximum Iteration	Gradient Descent	Stochastic GD	Regularization
49.22	0.0004	0.005	50000	FALSE	FALSE	FALSE
1.42	0.0004	0.005	50000	TRUE	FALSE	FALSE
24.16	0.0004	0.005	50000	FALSE	TRUE	FALSE
189.13	0.0004	0.005	50000	FALSE	TRUE	TRUE
49.22	0.04	0.05	5000	FALSE	FALSE	FALSE
4.42	0.04	0.05	5000	TRUE	FALSE	FALSE
115.57	0.04	0.05	5000	FALSE	TRUE	FALSE
177.20	0.04	0.05	5000	FALSE	TRUE	TRUE

III.II. Logistic Regression:

The primary idea behind Logistic regression is to establish a relationship between an input feature and a predictor variable which is often a categorical one for classification. There are different types of logistic regression models depending upon the number of classes the data is required to be segregated into:

Binary classification: This involves predicting if the sample falls into class 0 or 1. For example, an email is a spam or ham. The sigmoid function is used to calculate values for logistic regression involving two classes. Sigmoid pushes the input values to fall within the range [0,1].

$$Fx_i = 1/(1 + \exp{-x_i})$$

Multiclass logistic regression: It is used to classify records into more than two classes. A SoftMax function is used to calculate the probabilities. For a given sample we calculate the probability of it belonging to each of the n classes. The sum of all probabilities for each sample should add up to 1. We classify a record belonging to a particular class by selecting the one with the highest probability.

For our dataset, we have engineered a new column called **AQI_encoded** by using the below logic. This helps us to fit the data into a multinomial Logistic Regression model. The target classes are one-hot encoded before calculating the probabilities.

Categories for AQI classification	AQI_encoded values
Good	0
Moderate	1
Unhealthy for Sensitive Groups	2
Unhealthy	3
Very unhealthy	4
Hazardous	5

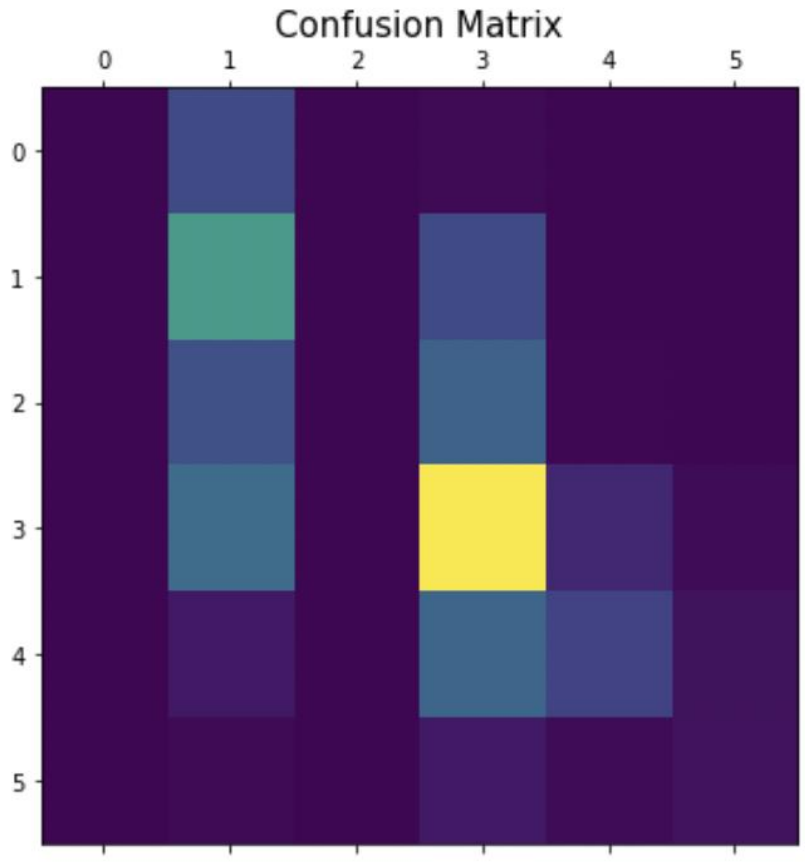
Results:

Below we show the results of the evaluation metrics:

	precision	recall	f1-score	support
Good	0.00	0.00	0.00	7882
Moderate	0.38	0.70	0.49	25614
Unhealthy for Sensitive Groups	0.00	0.00	0.00	18919
Unhealthy	0.51	0.68	0.58	48576
Very unhealthy	0.62	0.32	0.42	20677
Hazardous	0.43	0.32	0.36	4360
accuracy			0.47	126028
macro avg	0.32	0.34	0.31	126028
weighted avg	0.39	0.47	0.41	126028

Training Data Batch size: 1000			
Iterations	**Regularization**	**Learning Rate**	**Accuracy**
100	0.001	0.1e-3	43.218173739169075
1000	0.001	0.1e-3	44.95588281969086
1000	0.0001	0.1e-3	46.07150791887517
1000	0.0001	0.1e-5	46.78642841273368
5000	0.0001	0.1e-5	46.659472498175006
10000	0.0001	0.1e-5	42.843653791221

Training Data Batch size: 10000			
Iterations	**Regularization**	**Learning Rate**	**Accuracy**
100	0.1	0.1e-3	45.40578284190815
1000	0.01	0.1e-5	45.58590154568826
1000	0.0001	0.1e-3	45.77871584092424
1000	0.0001	0.1e-5	45.48830418637128
5000	0.0001	0.1e-5	45.98343225315009
10000	0.0001	0.1e-5	45.631129590249785




```
Performance metrics:
Accuracy: 46.78642841273368
```

III.III. Naive Bayes:

The core of a Naive Bayes model is prediction, diagnosis and reasoning. Variables generated randomly are represented by nodes and conditional probability is the information between these nodes. We combine conditional probability with prior probability to obtain a posteriori probability in order to achieve the effect of prediction. It is a simple model widely used due to its speed and faster learning rate. Bayes theorem helps us calculate the Probability of A happening provided B has occurred with the assumption of independence.

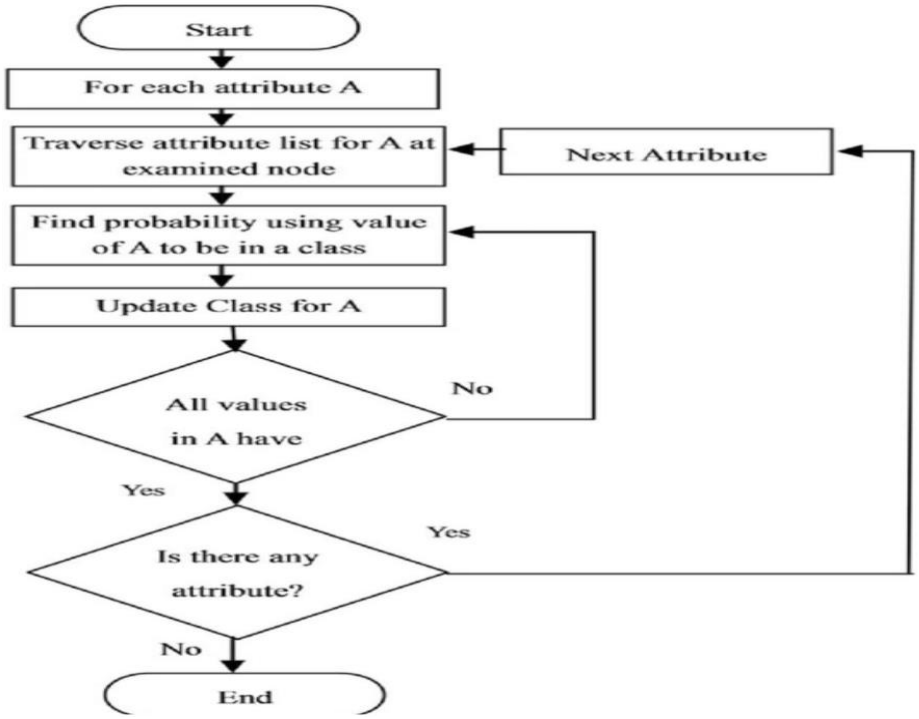


Fig 1.3: Flow Chart Showing End-End Naive Bayes Model Cycle

The following formula is used for calculate probability:

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_{i=1}^{n} P(B_i)P(A|B_i)}$$

The Probability of event B_i occurring such that event A has occurred is given by the above formulae.

The Gaussian Naive Bayes functions are typically derived from probabilities of the lower-order estimates from the training data. These may easily be updated as new training data come in. Because you just need to estimate the mean and standard deviation from your training data, the Gaussian (or Normal) distribution is the simplest to deal with. The assumption that the continuous values associated with each class are distributed according to a normal (or Gaussian) distribution is frequently made when working with continuous data. The characteristics and likelihood is predicated on:

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

Assuming variance is,

- is independent of Y (i.e., σi),
- or independent of Xi (i.e., σk)
- or both (i.e., σ)

The AQI classification categories were encoded into the below values in order to classify our predictions:

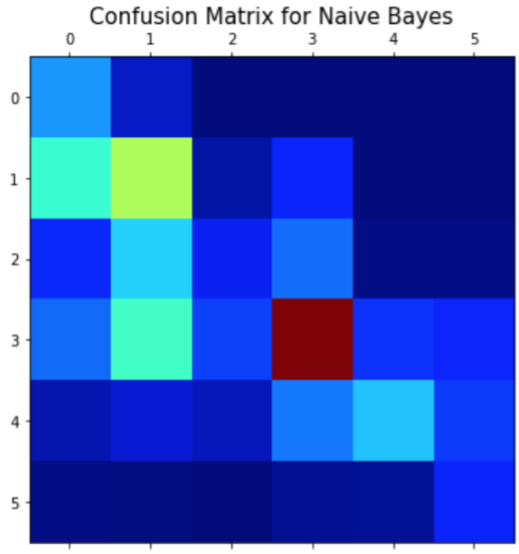
Categories for AQI classification	AQI_encoded values
Good	0
Moderate	1
Unhealthy for Sensitive Groups	2
Unhealthy	3
Very unhealthy	4
Hazardous	5

Results:

Below are the results obtained by applying batch size to data split while running the model:

Batch Size	Accuracy	Precision Value					
		Class					
		0	1	2	3	4	5
20%	43.18	24.86	38.37	27.24	61.77	61.87	28.51
25%	43.18	24.68	38.4	27.51	61.79	61.93	28.60
30%	43.28	24.9	38.42	27.7	61.94	61.62	28.33
35%	43.29	24.94	38.4	27.64	61.86	61.62	28.23

Batch Size	Accuracy	Recall Value					
		Class					
		0	1	2	3	4	5
20%	43.18	79.58	49.66	12.71	47.31	34.78	65.49
25%	43.18	79.64	49.54	12.64	47.34	34.9	65.37
30%	43.28	79.58	49.75	12.82	47.49	34.76	65.11
35%	43.29	79.34	49.97	12.78	47.49	34.78	64.95

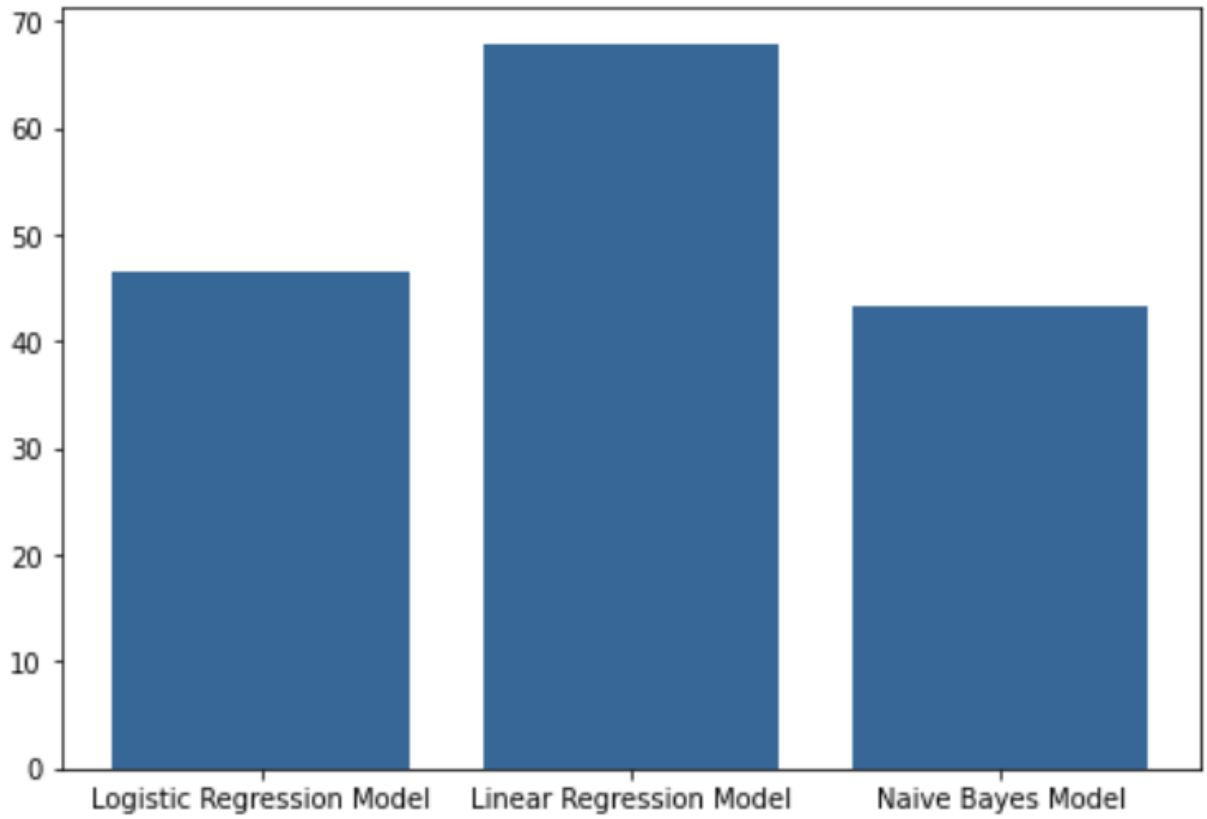



```
Performance metrics:
Accuracy:  0.4328892153462148
Precision:  [0.24940854 0.384752   0.27640604 0.61864231 0.61622968 0.28230361]
Recall:  [0.79349842 0.49971584 0.12786692 0.47492318 0.34788721 0.64953731]
```

IV. Performance Evaluation:

Below is a table showing model performance over the dataset:

Models	Accuracy
Linear Regression	68%
Logistic Regression	46.78%
Naïve Bayes	43.29%



V. Conclusion:

The air quality prediction model based on machine learning is used to predict the air quality in Beijing. The aim of the project was to research and study the various Machine Learning models available in order to implement the best fit for our dataset. As a team, we had the opportunity to clean the data, observe the importance of pollutants and their impact on the quality of air and implement ML algorithms to predict the AQI and observe how the predicted value differs from the original value. The following factors are used for the evaluation of the model, Temp, Pressure, SO2, NO2, O3, CO, PM10 and PM2.5 and the calculated AQI value is used as the target. The model is trained and verified by using the historical data on air quality to obtain comprehensive accuracy. The results show that the Linear Regression model predicted the air quality with a 68% accuracy, followed by Logistic Regression at 46.78% and Naive Bayes at 43.29%. In addition, wind, precipitation and other environmental conditions also play a factor that directly affects the quality of air. Therefore, taking these factors would help further improve the accuracy of our models and is one of the future research directions.

VI. References:

https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data
http://www.ijsei.com/papers/ijsei-67117-18.pdf
https://iopscience.iop.org/article/10.1088/1742-6596/2010/1/012011/pdf
BC. Liu, et al, "Urban air quality forecasting based on multi-dimensional collaborative Support Vector Regression (SVR): A case study of Beijing-TianjinShijiazhuang", PLOS, 2017